浅析互联网大数据在媒体业务的支撑应用

——以人民公安报社舆情监测系统为例

摘　要：为适应新形势下网络安全工作需要，网络舆情监测系统应运而生。该系统旨在对网络各类信息进行汇集、分类、整合、筛选等技术处理，实现最大社会效益，促进行业健康发展，维护社会稳定。本文通过梳理网络安全相关政策，分析网络舆情监测系统的应用，总结技术运用带来的社会效益等方法，以期给大家启发。

关键词：人民公安报；舆情监测；互联网；大数据；媒体融合
中图分类号：TP393　　文献标识码：A
文章编号：1671-0134（2019）06-080-03　　DOI：10.19483/j.cnki.11-4653/n.2019.06.022

文 / 魏春光

随着市场竞争的日益加剧，如何开发信息资源、利用信息资源，并实现信息资源的最大利益化显得尤为重要，越来越多的公众已意识到信息是一种潜在的生产力。根据中国互联网络信息中心（CNNIC）发布的第 43 次《中国互联网络发展状况统计报告》显示，截至 2018 年 12 月，我国网民规模达 8.29 亿，普及率达 59.6%。[1] 中国互联网已经进入全新发展阶段，各行各业也随之经历了不同程度的变革。

在互联网时代下，谁重视信息安全谁就发展稳定，谁重视网络舆情监测谁就会实现更大社会效益。因此，本文试图借助人民公安报社舆情监测系统分析监测信息的必要性，从而论证网络舆情监测对行业、对公众、对社会具有重要意义。

1. 政策梳理

党的十八大以来，以习近平同志为核心的党中央高度重视网络安全和信息化工作，紧紧围绕我国经济社会发展的总要求和大趋势，着力推动我国网络安全和信息化工作实现新发展，维护国家和人民安全。

2016 年 4 月 19 日，在网络安全和信息化工作座谈会上，习近平总书记指出，要树立正确的网络安全观，加快构建关键信息基础设施安全保障体系，全天候全方位感知网络安全态势，增强网络安全防御能力和威慑能力。同时，习近平强调，维护网络安全是全社会的共同责任，需要政府、企业、社会组织、广大网民共同参与，共筑网络安全防线。[2]

2016 年 8 月，国务院办公厅印发《关于在政务公开工作中进一步做好政务舆情回应的通知》。该《通知》指出，随着互联网的迅猛发展，新型传播方式不断涌现，政府的施政环境发生深刻变化，舆情事件频发多发，加强政务公开、做好政务舆情回应日益成为政府提升治理能力的内在要求。[3]

2018 年 4 月，国务院办公厅印发《2018 年政务公开工作要点》提出，增强舆情风险防控意识，密切监测收集苗头性舆情，特别是涉及经济社会重大政策、影响党和政府公信力、冲击道德底线等方面的政务舆情，做到及时预警、科学研判、妥善处置、有效回应。[4]

当今时代，网信事业正逐渐成为重塑国际经济、政治、文化、社会、生态、军事发展新格局的主导力量。网络安全关系着国家安危，可以说是“没有网络安全就没有国家安全”，确保网络安全成为确保国家安全的重要任务。

2. 大数据技术

人民公安报社舆情监测系统作为报社融媒体发展战略的核心系统之一，将实现对全网的舆情监控和分析，有效引导社会热点和公众舆论，有力发挥中央媒体优势，为国家公安事业发展服务。系统的建设将基于全媒体的舆情监测网络和分析机制，利用大数据等信息技术，科学、全面、高效地掌握网络舆情，对指定范围内的网站信息发布进行全面掌控，实现集“新闻、论坛、博客、微博、新闻客户端等网络信息实时监控，舆情信息传播渠道跟踪，溯源和舆情导控指挥”三大功能为一体的舆情监控分析平台。最终形成和生产出具有鲜明行业特色的舆情监测常规产品，包括行业的日、月、年度报告。

人民公安报社舆情监测系统利用当前最先进的分布式计算技术、数据管理与检索技术、数据智能分析技术，采用传统关系数据库、分布式数据仓库、分布式文件数据库相结合的方案，实现智能检索和数据高效管理，深度挖掘和智能分析数据，提供数据自动分类、自动聚类、自动关联、自动标引等一系列的智能分析，使数据得到最广泛的关联，进而挖掘知识。该舆情监测系统的建设目标是对互联网上媒体反映的舆论与民意实现全面有效的采集、分析、研判和表达，并及时有效响应。

2.1 平台特点

大数据是一种数据集合，其具有大容量、高精度和快速高效等特征。[4] 本项目以实现对报社关注的互联网

相关信息采集、专题事件分析、社会热点发现、重点内容监测、数据统计分析、舆情简报制作、检索、管理等功能为目标，力求达到内容全面、功能齐备、方便易用、开放兼容、安全可靠。总之，对舆情的全面了解与掌握是舆情监测系统的重要工作之一。

本系统的建设目的是：掌握网民主要观点和视角态度；了解媒体的报道情况和关注重点；自动生成舆情简报，及时响应突发事件，提高工作效率；能对特定事件进行持续性跟踪和分析；形成统一的运营服务平台，作为新闻选题采编工作的业务支撑辅助平台。

2.1.1 系统建设的必要性

建设舆情管理系统，首先是确保国家长治久安的需要，有利于建设好、利用好、管理好互联网，有利于维护改革发展的大局，有利于巩固党的执政基础。

其次是推动建立正确舆论导向是前提，有利于密切联系群众，及时准确掌握社情民意，有效引导网上舆论，把握舆情发展走向。

再次是适应未来网络舆论管理的迫切需要，有利于及时应对各类网络新媒体和移动互联网等媒介融合趋势，提高处理互联网舆情问题的准确性。

还有是运用高新技术手段是提升管理工作能力的迫切需要，有利于适应新时期信息化发展战略，完善电子政务体系。

最后是适应网络宣传工作与时俱进、创新发展的需要，有利于提高互联网从业人员管理能力和思想水平，发挥网络媒体的行业自律机制。

2.1.2 详细科学技术内容

（1）分布式数据管理技术：海贝大数据管理系统（Hybase）以存储、检索和统计为核心，采用弹性扩展架构设计的新一代大数据管理系统，它融合了全文检索、自然语言处理、索引分片、多副本机制、对等节点机制（去中心化）、列存储、内存索引等多项先进技术，为各类非结构化大数据分析应用提供非结构化大数据高效管理和智能检索。其具备以下优势：

扁平化设计：扁平化架构使单个节点故障不会影响整个系统对外提供服务；同时，该架构使系统具有良好的扩展性，可在线增加新的节点，扩展系统容量和增加对外服务能力。

异常感知可以自动恢复：当系统自动感知服务器处于异常状态时，可以进行自我修复。该系统是可以将硬件异常作为常见异常来处理的，不会因单个节点的异常导致整个系统不可使用。

柔性多引擎技术：该系统通过定义一个标准的引擎接口，采用多引擎机制。对于不同的应用需求，可使用不同的引擎，用户甚至可以自己构建引擎来扩展系统的数据处理能力。

支持异构数据：该系统支持结构化、半结构化、非结构化数据的统一检索。

高效分区索引机制：根据查询特点，该系统可将数据自动分区索引。

混合索引方式：该系统提供按词、按字、字词混合索引方式，满足不同应用场景对查全和查准的不同需求。

内存表：该系统支持在内存中建立数据表，适应数据量较少，但查询并发与响应速度要求很高的应用需求。

列存储：该系统支持列存储，实现特定数据列的高效访问，提高特定数据列的分类统计和排序的速度。

异步检索：支持异步检索模式，适应大开发（高连接数）的应用场景要求，避免了同步检索模式时消耗太多线程资源的问题。

多层次、多粒度的分布式 CACHE：该系统具有单节点的检索缓存和合并后的整体检索缓存，可以大大提高缓存命中率，减轻高并发下的检索节点压力，从而大幅度提高系统在高并发情况下的数据检索能力。

可扩展的检索模式：同根词检索、算法和词典结合的英文词根检索，准确率达到 99.9%。同时，支持基于同义词、主题词的扩展检索。

兼容 Hadoop 标准：TRSHyBase 和 Haboop 无缝集成，可以充分利用 HDFS 的可靠性，承担图像、音视频等大对象的存储。

（2）互联网信息采集：海量互联网数据实时监测，数据范围涵盖新闻、纸媒、论坛、博客、微博、微信、APP、搜索引擎等。舆情信息传播渠道跟踪，溯源和舆情导控指挥三大功能为一体的舆情监控分析平台。最终形成和生产出具有鲜明特色的舆情监测常规产品，包括各行业的日、月、年度报告。

（3）互联网信息智能处理：针对不同类型的舆情内容，Hybase 大数据管理系统利用先进的统计技术和智能文本分析挖掘技术实现数据过滤。该系统具有多语种识别和自动转码、自动分词、自动分类、自动聚类、自动热点发现、相似检索、文章排重、自动摘要、重点信息抽取等功能，可以根据实际工作需要，为舆情监控平台各项功能进行基础数据加工。

（4）全文检索功能：该系统可以按来源、时间、境内、信息源等多种分类检索，提供智能分析的信息检索服务。同时，不同用户，根据其权限检索相关的内容。如可对正文、标题、时间、作者、网站等进行高级检索，检索响应速度平均不超过 5 秒。此外，系统对用户可设置权限进行相关内容的检索。

（5）互联网信息分析应用：系统实现对重点信息的预警提醒，重点事件的趋势分析、网站分析、人物分析、热点分析，及自动生成舆情报告功能。权限上提供了完善用户和权限管理机制，充分保证情报信息内容的安全性。用户分组、分类，权限分级。系统支持按照分类进行权限控制，可控制用户也可控制角色，提供系统数据的安全性及应用性。提供多用户登录功能，对用户功能权限、关键词、栏目、专题、信息提供层级化管理设定。对文章进行管理，如置顶、收藏、隐藏、录入、编辑、审核，能对网页痕迹进行证据保留，并且利用探针功能发现原文连接是否有效。系统提供完整详细的日志，根据日志能够获得用户的登录和管理情况；日志能够根据

条件进行查询，实现系统操作日志的详细记录及各部门、各用户的应用统计信息，方便审计管理员进行应用审计。

2.2 技术创新点：大数据管理

网络系统逐渐复杂化，这是技术应用与发展的趋势，随着数据量的持续增长，信息正在实现由TB级到PB级的跨越式前进，使数据分析的纬度指标变得更加广泛。[6]针对本项目研发的大数据管理系统，一方面可以实现结构化数据、半结构化数据、非结构化数据的统一管理和检索；另一方面，还顺应了“非结构化数据的结构化处理、结构化数据的非结构化处理”的技术趋势。

2.2.1 信息采集技术

本项目在采集方面的关注重点是搜索引擎技术很少涉及的深层次采集技术（面向DeepWeb）。网络应用技术快速发展，网络信息呈现出一定的“异构”特点。随着互联网社区化的发展、Web2.0的崛起，以HTTP为网络传输协议，以HTML为展示格式的网络信息已不能适应发展所需，网页所蕴含的内容正发生着深刻的变化。原来以网站/网页内容为主导的互联网，逐渐演变为网站、微博、微信、论坛（社区）、博客等信息共存的局面。微博、微信、论坛、博客等平台上蕴含着大量的信息，已然成为互联网上信息的重要来源，而且对行业搜索引擎建设来说，这些平台上的信息比普通网站上的信息具有更重要的使用价值。

系统不仅对数据进行智能分析及挖掘，还需在此基础上充分利用数据智能分析技术获取的知识标签，对知识进行融合、加工，进而构建知识图谱，使用户能够像使用百科全书一样查询、浏览知识词条，以及具备广泛关联关系的知识图谱。系统需充分利用文本挖掘获取的元数据内容创建“故事流”式的服务，为新闻生产提供智能辅助。系统需从正负面信息、关注程度、传播速度等方面对传播内容进行传播分析，获取传播效果，为报社智能决策奠定基础。

综上所述，舆情监测系统具备承上启下、兼容并包的作用，既可以满足系统建设的功能需求，又能盘活新增的海量数据资产，实现数据的增值及再利用，为报社的新闻发现和智能创作支撑，进而促进媒体融合发展，切实贯彻落实习近平总书记在党的新闻舆论工作座谈会上的重要讲话精神。

2.2.2 与当前国内外同类研究、同类技术的综合比较

分布式大数据管理系统：实现海量数据的组织和管理需要一个可扩展的存储和处理框架。目前，采用廉价计算机的极具扩展性的分布式云计算环境不仅引起了商业巨头IBM、EMC、微软等公司的重视，而且在Google、Amazon、Yahoo等公司已经取得成功。云计算环境一般包括可扩展的文件系统、并发处理的操作原语和可靠的数据存储。由于对海量数据的管理需要采用全新的计算模式和存储模式，因此，业界如Google、Yahoo、微软和IBM等企业和科研机构充分利用底层云计算环境所提供的数据存储和并发处理的功能实现海量数据的存储和管理。

分布式计算环境的发展为海量数据提供了存储和处理基础。各大公司开始构建分布式计算环境的基于SOA的海量数据集成系统。从目前进展情况看，存在的主要问题包括：目前的非结构化数据中的元数据可能包括锚文字、日期等通用元数据，或者用户手工输入的信息，尚未有效集成信息提取和非结构化数据管理；分布式计算模型能够方便应用关键字查询，但是对数据条件查询并没有很好的优化，数据查询处理的效率有待提高；海量非结构化数据系统的Pay-as-you-go的方式需要进一步支持，包括底层存储对不同属性合并、分解、优化存储等。

2.2.3 智能文本处理技术

国外开展文本挖掘和信息抽取等研究比较早，研究机构众多，比较著名的有：卡内基梅隆大学、马里兰大学、加州大学伯克利分校、IBM公司等。国内从80年代开始文本挖掘和信息抽取等研究，从事该领域研究的主要机构有：北京大学、清华大学、哈工大、中科院计算所、微软亚洲研究院等。我国对这方面的研究非常重视，国家863计划等多次组织了对分词、分类、摘要、关键词标引、信息抽取、褒贬分析等文本智能技术的专门评测，这些评测的举行极大地推动了国内的相关技术发展。

针对本项目研发的智能文本处理系统，利用先进的统计技术和智能文本分析挖掘技术针对不同类型的舆情实现数据内容过滤，多语种识别和自动转码、自动分词、自动分类、自动聚类、自动热点发现、相似检索、文章排重、自动摘要、重点信息抽取等功能，为舆情监控平台各项功能进行基础数据加工。

3. 大数据技术带来的社会效益

网络已经成为我国信息传递的主要方式，因此对网络环境必须十分重视，只有维护好网络环境才能够真正发挥网络的作用，更好地为用户带来便利，同时也对经济的发展和文化的发展带来积极的影响。[7]本项目是全面贯彻落实习近平总书记在党的新闻舆论工作座谈会上重要讲话精神的重要组成部分，是将大数据技术的研究成果应用于媒体转型实践的重要步骤，具有重要的社会效益。

3.1 全面贯彻落实习近平总书记在党的新闻舆论工作座谈会上重要讲话精神

本项目紧紧围绕习近平总书记重要讲话精神，坚持正确的政治方向和舆论导向，紧抓信息化发展的历史机遇，加速信息领域核心技术突破进程，维护网络社会安全，营造风清气正的网络空间，充分发挥信息技术对经济社会发展的引领作用。

3.2 有利于遏制有害信息及言论的传播扩散，以正确的舆论引导人

网络舆情可以了解社情民意，对网络民意的有效数据进行科学筛选、量化统计和分析，并根据实践经验，紧密结合历史发展和中国国情进行研判，对倾向性和苗头性问题有超前预测作用。[8]通过本项目的建设，可以加强对网络舆论态势的把握，做好舆情收集和综合研判，为中央决策提供参考；还可以搭建政府与群众间的“绿

（下转第117页）